Check for updates

General Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(4) 556–573
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245918806489
www.psychologicalscience.org/AMPPS

Peer-Review Guidelines Promoting Replicability and Transparency in Psychological Science

William E. Davis[1], Roger Giner-Sorolla[2], D. Stephen Lindsay[3], Jessica P. Lougheed[4], Matthew C. Makel[5], Matt E. Meier[6], Jessie Sun[7], Leigh Ann Vaughn[8], and John M. Zelenski[9]

[1]Department of Psychology, Wittenberg University; [2]School of Psychology, University of Kent; [3]Department of Psychology, University of Victoria; [4]Department of Human Development and Family Studies, Purdue University; [5]Talent Identification Program, Duke University; [6]Psychology Department, Western Carolina University; [7]Department of Psychology, University of California, Davis; [8]Department of Psychology, Ithaca College; and [9]Department of Psychology, Carleton University

Abstract

More and more psychological researchers have come to appreciate the perils of common but poorly justified research practices and are rethinking commonly held standards for evaluating research. As this methodological reform expresses itself in psychological research, peer reviewers of such work must also adapt their practices to remain relevant. Reviewers of journal submissions wield considerable power to promote methodological reform, and thereby contribute to the advancement of a more robust psychological literature. We describe concrete practices that reviewers can use to encourage transparency, intellectual humility, and more valid assessments of the methods and statistics reported in articles.

Keywords

replication, reproducibility, publication bias, *p*-hacking, validity, peer review

Received 3/12/18; Revision accepted 9/19/18

Psychological science is undergoing a "renaissance" (Nelson, Simmons, & Simonsohn, 2018) or "credibility revolution" (Vazire, 2018) in understanding of statistical inference, in standards for methodological rigor, and in expectations of what should be reported in scientific communications. These developments have come with a realization that previous standard practices, most notably the focus on multiple conceptual replications in a single research article, were not enough to ensure replicable and robust science. There is a growing call to raise the field's standards (Vazire, 2018), and this in turn will require access to more details of studies' methods, analyses, and data than was previously typically provided—information that is still often omitted from reports.

Our aim in this article is to provide recommendations for reviewers to promote transparency, statistical rigor, and intellectual humility in research publications. Well-informed peer reviewers help journal editors make better decisions not only about whether a piece of research should be published, but also about how the work is reported if it is published. Reviewers can influence reporting practices by requesting the transparency necessary for all readers to assess the quality of the evidence and the validity of conclusions (Morey et al., 2016; Vazire, 2017). Our advice applies particularly to quantitative research in psychology, but is also relevant to research in other fields of science, especially those that use inferential statistics.

Corresponding Author:
Roger Giner-Sorolla, School of Psychology, University of Kent, Keynes College, Canterbury CT2 7NP, United Kingdom
E-mail: rsg@kent.ac.uk

This article grew out of a workshop, "How to Promote Transparency and Replicability as a Reviewer," at the 2017 meeting of the Society for the Improvement of Psychological Science. Workshop participants (including this article's authors) read existing advice on reviewing provided for the occasion by 22 journal editors (see Lindsay, 2017, and Lindsay, Giner-Sorolla, & Sun, 2017), Roediger's (2007) "Twelve Tips for Reviewers," a chapter on reviewing by Tesser and Martin (2006), and an excerpt from Commitment to Research Transparency (Schönbrodt, Maier, Heene, & Zehetleitner, 2015). Workshop members then put together a set of new recommendations aimed at promoting transparency and replicability. In this article, we first explain some of the issues underlying our advice and then present our recommendations.

The New Approach to Statistical Inference and Reporting

Most empirical reports in psychology use null-hypothesis significance testing (NHST) as a metric of evidence. In NHST, inferential analyses such as *t* tests yield estimates of the probability (p) of the obtained result (or a more extreme result) occurring by chance under the null hypothesis of no effect. If p is low enough, usually under the conventional $p < .05$ threshold, the result is deemed "statistically significant." Significance can be taken as a heuristic indicating that the direction of the effect in the sample is likely to be the same as in the population (Krueger & Heck, in press). However, problematic practices call into question the usual ways in which statistical significance—in particular, the criterion of $p < .05$—has informed publication decisions.

NHST is accurate only in confirmatory research, in which the hypotheses to be tested and the method of testing are specified *before* the data are examined (Simmons, Nelson, & Simonsohn, 2011). But in practice, researchers sometimes decide which analyses to run on the basis of which tests produce the most favorable results, and then report those analyses as if they had been planned in advance. Similarly, researchers sometimes adjust their procedures while analyzing their data (e.g., dropping some subjects, observations, dependent variables, or conditions; adding covariates; transforming measures) and fail to report these adjustments. All these practices may reflect a desire for brevity and a stronger narrative—spurred as much by editorial standards as by the authors themselves.

This sort of flexible, post hoc approach to NHST has been common practice in many areas of psychology (John, Loewenstein, & Prelec, 2012). Unfortunately, these practices make p values misleading. Different critics have used different terms to highlight various aspects of the problem (e.g., HARKing—Kerr, 1998; researcher degrees of freedom—Simmons et al., 2011; p-hacking—Simmons, Nelson, & Simonsohn, 2012; the garden of forking paths—Gelman & Loken, 2013; questionable research practices—John et al., 2012). Regardless of terminology, these practices can exaggerate estimates of the sizes of effects and inflate the risk of falsely rejecting the null hypothesis. When "significant" p values obtained via undisclosed flexibility are presented as if they arose from planned tests of hypotheses, readers are likely to conclude that the evidence is stronger than it actually is.

It is good and proper for researchers to conduct exploratory research as well as hypothesis-testing research. Poking around in one's data, speculating about unexpected patterns, is a great way to generate ideas. For conducting such exploratory analyses, confidence intervals and estimates of effect size are useful tools (e.g., McIntosh, 2017). But NHST p values become meaningless when the data drive decisions about which tests to run and how to run them, because more risks have been taken than the p value takes into account. At a minimum, reviewers and readers need to know how researchers made their data-analysis decisions.

Vazire (2017) drew an analogy between readers of science articles and used-car shoppers: Transparent reporting puts readers in a better position to tell the difference between "lemons" and trustworthy findings. One powerful tool for promoting such transparency is preregistering the research plan (see Lindsay, Simons, & Lilienfeld, 2016; van 't Veer & Giner-Sorolla, 2016). Preregistration makes clear which aspects of a study and its analyses were planned in advance of data collection. Openly sharing data and materials (e.g., tests, stimuli, programs), and explicitly declaring that methodological details have been completely reported (e.g., Simmons et al.'s, 2012, "21 word solution"), can also help readers to assess the evidence value of an empirical report.

To allow for correction of mistakes in reporting and for exploration of alternative analyses and explanations, transparency requires that researchers make their raw data available to other researchers, along with codebooks and analysis scripts. Despite protocols requiring such sharing for verification (e.g., Section 8.14 of the American Psychological Association's, 2017, ethical principles), the availability of data has often been poor (e.g., Wicherts, Borsboom, Kats, & Molenaar, 2006). Finally, authors can also advance transparency by providing more comprehensive descriptive statistics, such as data graphs that show the distribution of scores. Making defensible claims in research reports also entails intellectual humility about the limitations of one's own perspective and findings (Samuelson et al., 2015).

Scientific claims require a realistic perspective on the generalizability of one's own research and views. In moving from a standard that prioritizes novelty to one that emphasizes robustness of evidence, claims about the importance of any one study or series of studies should be limited, and replications should be encouraged. Researchers should also strive to be aware of the assumptions they bring to conducting and evaluating research—for example, their ideas about what constitutes a "standard" or "unusual" sample (see Henrich, Heine, & Norenzayan, 2010) or their preconceptions about research that has political implications (Duarte et al., 2015).

Over the past decade, some journals in psychology and other fields have adopted more open reporting requirements, such as those outlined in the Transparency and Openness Promotion (TOP) guidelines (Center for Open Science, n.d.-b; Nosek et al., 2015). More than 5,000 journals and organizations have become signatories of the TOP guidelines, and more than 850 journals have implemented the standards. However, many journals have not changed their policies, and editors and reviewers vary in implementing these reforms. Our aim with the following recommendations is to provide concrete guidelines showing how you, as a peer reviewer of empirical research articles, can encourage transparency, statistical rigor, and intellectual humility. We organize these guidelines, roughly, in the order they will come up as you deal with a review. Appendix A gives a slightly reorganized outline of our advice that can be used as a checklist during the review process.

Preparing to Review: Know Your Stuff

To understand and communicate criticism of research you review, you need to have a solid grasp of the key statistical issues. Appendix B lists selected educational resources, and we discuss some of these issues in the next section. Although specific statistics applications vary across fields, you should sharpen your understanding of the following concepts that often are forgotten after postgraduate statistical training:

- The logic of NHST: If you understand why the p value is not itself "the probability that the null hypothesis is true" (e.g., Cohen, 1994), you have come farther than many people.
- The need for a priori specification of hypothesis tests: In addition, it is important to know about methods used to control selective reporting, such as preregistering experiments; reporting all analyses, including those that might be labeled as exploratory and post hoc; providing methodological disclosure statements (Simmons et al., 2012); and openly sharing materials.
- Assumptions underlying frequently used statistical tests in your research area: In particular, it is important to know when a given test is not robust to violations.

One source of inspiration is the American Psychological Association's Journal Article Reporting Standards (JARS; Appelbaum et al., 2018). These guidelines list desirable features for reporting in all types of research articles, including those involving qualitative, meta-analytic, and mixed methods. Using JARS as a checklist, you can look for the methodological and statistical considerations that are particularly important to report in your area of research and carry out further reading to ensure that you understand their rationales.

Reading and Evaluating the Manuscript

Evaluate statistical logic and reporting

You might think that all editors of scientific journals in psychology are statistically savvy, but you would be wrong. Unfortunately, it is possible to become an eminent scholar and gatekeeper in psychology while keeping one's statistical knowledge focused on the skills that help get articles published, rather than on best statistical practices. Even if journals espouse improved statistical standards or refer authors to general guidelines, such as those in the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010), editors do not always enforce such guidelines before sending manuscripts to reviewers. It is often up to you, the reviewer, to insist on complete statistical reporting for the sake of transparency.

Of course, editors and authors may privilege other goals, such as manuscript readability or word-count limits, above full statistical reporting. Your suggestions for increasing the amount of reporting should take into account what is possible at the journal, as specified in its submission guidelines (sometimes known as "Guide for Authors" or "Instructions for Authors"), which should be available on the journal's Web site. Limitations caused by restricted word counts, for example, can be overcome by adding details in supplementary online materials (which many journals now offer) or on public repositories, such as the Open Science Framework (http://osf.io).

Beyond the journal's standards, the issues you look for will depend on your own knowledge and preparation. Here are several frequently encountered issues:

- Many psychology studies cannot obtain precise results because their sample sizes are not sufficient to provide accuracy in parameter estimation (AIPE; Maxwell, Kelley, & Rausch, 2008; see also Cumming, 2014). That fact has been known for decades (Cohen, 1962), but only recently has awareness of it become widespread. Accuracy allows inference to go beyond a merely directional finding, allowing comparison of the observed effect size with effect sizes for other known influences on the outcome and evaluation of the finding as a potential basis for real-world applications. Precision for planning (Cumming & Calin-Jageman, 2017), AIPE, and statistical power analysis can all help readers judge the sensitivity of methods, which has implications for interpreting both positive and null results. Rather than criticizing a study on the basis of your idea of what a "low *N*" looks like, it is preferable to use any of these techniques. Some methods, such as repeated measures designs, can yield precise results or high power with a surprisingly small number of subjects (Smith & Little, 2018).
- Effect sizes, and related statistics such as confidence intervals, are important adjuncts to significance tests that help readers interpret data more fully, especially when samples are unusually large or small (Cumming, 2014; Howell, 2010). Even if effect sizes are reported in Results sections, check to see that the discussion of results takes into account their magnitude and precision, and that conclusions are not based only on *p* values.
- Power analysis tests the likelihood of rejecting the null hypothesis if the alternative hypothesis is true, and journals are increasingly requiring that such analyses be reported. Not all power analyses are equal, though. Post hoc power analyses, for instance, are uninformative, being merely a function of the *p* value (Goodman & Berlin, 1994). Best practice is to base the sample size on a reported a priori power analysis and to include a rationale for deciding the expected effect size that was input to these calculations (e.g., prior literature, estimates of the typical effect size for the field and methodology if an entirely new effect is being studied). If power analysis was not done a priori, you can still request a sensitivity power analysis that takes the actual *N* and a desired level of power, and outputs the minimum effect size that the study could have detected (Faul, Erdfelder, Lang, & Buchner, 2007). A study that can detect only a conventionally large effect at 80% power is not well powered to detect the small- and medium-sized effects that are more characteristic of many areas of psychological research. For reasons to prefer well-powered research, see the section titled Evaluate Sensitivity as Well as Validity.
- *Optional stopping* refers to the practice of using the outcome of a hypothesis test on preliminary data to decide whether to stop or extend data collection. Researchers might plan to stop data collection after a certain number of cases if the hypothesized effect is then statistically significant, and to continue data collection if it is not. This procedure might continue until the criterion significance is reached or until a maximum number of cases has been reached. Optional stopping can be acceptable if the researcher adjusts the alpha level accordingly (e.g., Lakens, 2014a; Sagarin, Amber, & Lee, 2014) or uses appropriate Bayesian analyses (e.g., Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). However, using the unadjusted .05 threshold with optional stopping inflates the Type I error rate. It is hard for a reviewer to detect unreported stopping rules, but you can look for or request a disclosure statement that explicitly describes how sample size was determined at each stage (Simmons et al., 2012).
- Descriptive statistics, such as cell *n*s, means, standard deviations, and correlations between multiple measures, are sometimes omitted from advanced statistical reports. Insist on seeing them anyway, because they may reveal underlying problems that qualify the fancier analysis. For example, means might be very low or high on the scale and low in variance (an indication of a floor or ceiling effect), and thus violate the assumptions of the statistical test; or two variables might be so highly correlated (e.g., .8 or above) that drawing distinctions between them is problematic. And if a complex, multivariable model gives results that appear at odds with the basic zero-order correlations in the data, it is important to understand why.
- Basic statistical errors are surprisingly common in published research (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Being roughly familiar with the formulas for degrees of freedom in commonly used statistical tests (e.g., Howell, 2010) can help you detect discrepancies between reported subject numbers and the actual numbers tested. There are also tools for checking whether the figures after the decimal point in a reported mean are impossible to obtain given the reported *n* in a condition (e.g., a mean of 2.50 with an odd number of data points; Brown & Heathers, 2016). Both problems may point to

undisclosed missing or excluded data. You might also want to run statcheck (Epskamp & Nuijten, 2016; Nuijten, Epskamp, & Rife, 2018) on manuscripts you review. This free program detects discrepancies among some of the most common inferential statistical indices (e.g., *F, r, t, z*), the reported degrees of freedom, and the reported *p* values.

Assess any preregistrations

To increase the appearance of confidence in results, it has been a common practice in psychology to report the outcome of exploratory analyses as though they had been planned a priori (John et al., 2012). Preregistration involves posting a time-stamped record of method and analysis plans online prior to data collection. It is intended to make analytic flexibility transparent, helping reviewers better evaluate the research. A common misconception is that a preregistration is meant to restrict the analyses that are performed; actually, preregistration does allow additional post hoc analyses, but the purpose of preregistration is to make sure that post hoc analyses are clearly labeled as such (e.g., van 't Veer & Giner-Sorolla, 2016).

If a preregistered plan for the research is available, it is important to assess the level of completeness and detail in that plan compared with the procedures reported in the article. Some "preregistrations" are so brief and vague that they do little to identify when post hoc liberties have been taken, providing only the illusion of transparency. Norms for assessing the quality of preregistrations are still in development (for one protocol, see Veldkamp, 2017). If researchers deviated substantially from their preregistered analyses, even for good reasons (e.g., the data failed to meet assumptions of the proposed test), you can ask them to also report the outcome of the preregistered analyses (e.g., in an appendix) for full transparency.

If the research under review was not preregistered, it may be difficult to tell which analyses were planned in advance and which were data dependent, but some clues may lead you to suspect post hoc analysis. For example, data exclusion rules or transformations might be reported only in the Results sections and without any explicit rationale, or may vary from one study to the next without justification. The concern here is that the researchers may have (not necessarily intentionally) made analytic decisions to produce significant results that would not be replicated if alternative reasonable analytic specifications were used or if a new data set were analyzed. That does not mean that those results have no value, but they should be viewed with skepticism pending direct replication.

You can ask researchers to address concerns about post hoc flexibility in your review. The strongest reassurance would come from a direct, preregistered replication. However, you can also ask the authors to indicate which analyses, if any, were exploratory or to adopt a more stringent standard for statistical significance (e.g., $p < .005$; Benjamin et al., 2018). Finally, you can ask the researchers to demonstrate that their findings are robust under reasonable alternative specifications (e.g., when included covariates are omitted, different exclusion criteria for subjects are used, or different model specifications are used; see Simonsohn, Simmons, & Nelson, 2016; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

Check data and materials

If the authors submitted data, materials, or analysis code as part of the review process, or if they provided a link to a preregistration document detailing their data-collection and analysis plans, you should determine whether these resources are in a usable form. If the data and materials are not available or usable, let the editor know and ask if there is a way to obtain them. When they are available, we encourage you to examine them for completeness and accuracy. Variables in the data set should clearly correspond to the variables reported in the text. Materials should allow a third party to rerun the study, and should map clearly onto the conditions, variables, and reporting. Running analyses with available data is usually beyond the call of a reviewer's duty, but might be worth the effort if it is helpful for checking apparent errors or identifying strong alternatives to the authors' conclusions.

Go beyond "p < .05 per study"

For a long time, in many areas of psychology, reviewers judged whether a study supported a hypothesis by whether its key test was significant at $p < .05$. A multistudy report was judged to support its hypothesis only when each study's key result was significant. To meet these standards, authors often omitted (or were asked to omit) studies with nonsignificant results, even though statistically they were consistent with evidence favoring the hypothesis. Another part of playing this game was p-*hacking*: selectively stopping data collection, excluding observations or conditions, applying data transformations, exploring covariates, or reporting one analysis out of many in order to achieve $p < .05$ (Simmons et al., 2011).

The distribution of *p* values from all tests of a true hypothesis should include relatively few results between $p = .01$ and $p = .05$ (Simonsohn, Nelson, & Simmons,

2014). The higher the statistical power of the tests, the larger the proportion of results with $p < .01$, and the fewer nonsignificant results (assuming a true effect). For example, if power is 80%, then about 59% of confirmatory tests should yield *p*s below .01, whereas only about 21% should yield *p*s between .01 and .05 (Lakens, 2014b; see also Magnusson's, 2018a, interactive calculator at http://rpsychologist.com/d3/pdist/). But some literatures in psychology report too many significant results relative to the power of the studies (Francis, 2014; Schimmack, 2012). And although authors disagree on the evidence for a "bump" in reported *p* values just under .05 (Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016; Masicampo & Lalande, 2012), there is a growing awareness that .05 is not a hard cutoff, and that single values close to it on either side are weak evidence (see Benjamin et al., 2018, and Lakens et al., 2018, for contrasting views on whether or not psychology should set alpha at .005.)

So, be wary of multiple studies, each with the key *p* value just under .05. Values in this range are infrequent enough, and it should be even more rare to see them across multiple studies. The pattern might have arisen by chance, but you should seek assurance that it is not due to selective reporting or *p*-hacking. A detailed and accurate preregistered analysis plan provides the greatest confidence (Lindsay et al., 2016; van 't Veer & Giner-Sorolla, 2016). Without such evidence of constraints on researcher degrees of freedom, you might look for or request a disclosure statement indicating that all measures, manipulations, and exclusions are reported (Simmons et al., 2012; see Nosek et al., 2017, for a standard reviewer statement).

Inzlicht (2015) gave an account of a lab that was encouraged to report all studies it had run to test a hypothesis, instead of just the significant ones, precisely because a manuscript it had submitted showed a pattern of *p* values unusually close to the significance criterion. When the lab's "file drawer" of nonsignificant findings was included, the overall picture still supported the hypothesis, albeit with a more modest effect size. Reporting all relevant studies, excluding only ones that fail methodological checks independently of the hypothesis, is a practice in line with both commonsense reporting ethics and the standards of professional bodies (American Psychological Association, 2010, p. 12). Although it is sometimes difficult to know when an additional unpublished study is part of the same or a different line of research, reviewers should encourage full reporting of studies that would have reasonably been included to support the argument of the manuscript at hand had they yielded significant results.

Reviewers should also place less emphasis on the *p* values of single studies than has typically been the case in the past. Better evidence can be gained from measures of precision (e.g., confidence intervals, credible intervals) or Bayes factors, which provide a symmetrical measure of evidence for the null and alternative hypotheses (Cumming, 2014; Wagenmakers et al., 2018). Often, when a series of studies is presented, better understanding can be had by aggregating comparable results over that series rather than commenting on each study's significance individually (Goh, Hall, & Rosenthal, 2016),

Aggregate evidence, however, becomes unreliable if only significant studies are reported. To mitigate publication bias, you can ask for an internal meta-analysis of all relevant studies conducted by the research team, which may include studies that were not included in the original report. But, by the same token, you should have realistic expectations about what a fully reported set of tests of a true hypothesis looks like (Lakens & Etz, 2017). Even if the proposition is strongly supported, this set can sometimes include nonsignificant results here and there.

Also, these considerations should not stop you from recommending publication of methodologically strong single-study manuscripts. One high-powered study can be more informative than several underpowered studies (Schimmack, 2012).

Evaluate measurement validity

Reviewers should make sure that the constructs discussed in a manuscript were indeed the constructs that were measured in the project. Ideally, an assessment should be sensitive to the differences that the researchers intended to measure (Borsboom & Mellenbergh, 2004). The interpretation of findings based on improperly validated measures can be meaningless at worst, and is suspect at best. Accessible discussions of these issues can be found in Flake, Pek, and Hehman (2017) and Fried and Flake (2018). Questions relevant to the validity of measures include the following:

- Have the authors reported scale reliabilities computed from their data? Indicators of internal consistency, such as Cronbach's alpha, are important to include but are commonly misreported as indicators of validity (Flake et al., 2017). In particular, a high alpha does not speak clearly to whether constituent items represent a single dimension or multiple dimensions. Factor analysis is needed to assess whether item intercorrelations match the intended structure, one aspect of valid measurement.
- Did the authors use previously validated measures? Check for reporting of, or references to, validation studies of the measures, including tests for construct, convergent, and divergent validity.

- Did the authors use measures as originally developed and validated, or have they modified the original scales? Are any modifications well justified and fully reported? Modifying scales without reporting the full details can complicate replication studies, and making modifications without assessing the validity of the resulting scales can lead to uncertainty in measurement.
- Did the authors report findings based on single-item measures? Single-item measures may not adequately capture the intended constructs. They require special consideration and validation (see Flake et al., 2017).

If you find that answers to any of these questions are unclear, it is important to request the missing information in your review. Authors should be encouraged to address weaknesses in measurement validity in the Discussion section of their manuscript, where they can describe specifically how uncertainty in the measures used may affect the interpretation of the results and the generalizability of the study.

Evaluate sensitivity as well as validity

Measurement concerns are part of a larger issue that is becoming more important with increased understanding of methodology: sensitivity. Traditionally, psychology reviewers have been keen to point out alternative explanations for a significant, or *positive*, result. Confounded manipulations, conceptually ambiguous measures, and statistical artifacts are just a few things that can threaten the interpretation of apparently positive results. Certainly, reviewers should stay on the lookout for all such issues.

In contrast, psychology reviewers are often less attuned to problems that might compromise the interpretation of nonsignificant findings, such as small sample size, weak manipulations, poor measurement reliability, restricted range, and ceiling or floor effects. Such flaws can reduce a method's *sensitivity* (ability to detect a positive result). Low sensitivity may obscure a phenomenon that exists in the population but is missed or underestimated in the sample, clouding the interpretation of nonsignificant results and casting doubt on the replicability of significant results. A common misconception (criticized by Loken & Gelman, 2017) is that a positive result is all the more impressive for having "survived" a study with low sensitivity. Reviewers should reject this view and look out for flaws in the sensitivity as well as validity of methods.

Low sensitivity raises the likelihood that a significant result is a false positive, especially when the finding is unlikely (Ioannidis, 2005, 2008; Zöllner & Pritchard, 2007). For example, if a finding is only 10% likely to be true and statistical power is low (50%), then 47% of $p < .05$ results will reflect a false effect. The false-positive problem, then, is likely to be particularly pernicious for surprising, counterintuitive findings not well supported by theory.

Low-sensitivity methodology also sets a bad example. A lab that uses it is more likely than other labs to waste their effort on a false-negative finding, and their findings are less likely to be replicated. And in a climate of low-sensitivity methodology, selective reporting can be justified more readily. If a study did not work, it is easy to say that the methods must have been bad, rather than to take the results as evidence against the hypothesis (LeBel & Peters, 2011). Finally, many inferential statistical tests lose their robustness to violations of data assumptions when sensitivity is low (e.g., because of a small sample size).

In experimental research, a particularly relevant sensitivity issue is manipulation validity. It is common for researchers to take a shortcut and assume that an effect of an independent variable on a dependent variable is sufficient proof that a manipulation is valid. But this assumption conflates the effect being tested (does change in the independent variable relates to change in the dependent variable?) with the validity of the manipulation (does the manipulation effectively change the independent variable?). Especially when results are null, either in original research or in a subsequent replication, showing that the manipulation is valid in the sampled population can help rule out manipulation failure as a prosaic explanation.

Ideally, a manipulation will be validated on a criterion variable that directly measures the independent variable. For example, if thoughts about power are being manipulated to be more accessible, then power words in a decision task should be responded to more quickly in the experimental than in the control condition. This test of the manipulation might be done in the same study that tests the main hypothesis, as a manipulation check. If there are concerns about subjects' awareness of the manipulation, though, the testing can be done on a separate sample (Kidd, 1976). Although manipulation checks have previously been criticized as unnecessary (Sigall & Mills, 1998), such critiques were based on their inability to shed light on positive results. With an increased emphasis on publishing and evaluating null results, testing manipulations has become more important.

Know how to evaluate null claims

Nonsignificant p values do not, by themselves, provide evidence for the null hypothesis. Evaluate a conclusion that an effect is nonexistent as carefully as you would evaluate a claim that it exists. Values of p greater than .10 are often obtained when the null hypothesis is false but sensitivity is low. If a manipulation causes a half-standard-deviation change in the population mean of a dependent variable (i.e., $d = 0.5$), then about half of experiments comparing two independent groups of 23 subjects will fail to reject that false null hypothesis at the .05 level (i.e., statistical power is only 50%). Bayesian approaches provide a more useful way to assess how much the data favor the null hypothesis (Wagenmakers et al., 2018). Alternatively, equivalence tests based on NHST (Lakens, 2017) can be performed. Both procedures depend on assumptions about the range of functionally null effect sizes, which should be described before reporting the results of the procedures. One does not need to be an expert in Bayesian or equivalence statistics to request that authors do more to justify or qualify the conclusion that an effect is nonexistent.

The general problem of drawing misguided inferences from nonsignificant p values can crop up in many other forms. For example, if a nonsignificant chi-square statistic (or change in chi-square) in a model-fitting analysis is used as the basis for concluding that the model fits (or that two models fit equally well), you should consider whether the study was sufficiently powered to detect misspecifications (Hu & Bentler, 1998). Also, if researchers claim to find "full mediation" on the basis of a nonsignificant direct effect (setting aside more general issues with statistical mediation; Bullock, Green, & Ha, 2010), you should consider how much power they had to detect small direct effects. In both cases, you can ask researchers to provide power analyses or qualify their conclusions.

Moreover, the difference between significant and nonsignificant is often itself not statistically significant (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Be especially wary if authors interpret a significant effect in one condition or experiment versus a nonsignificant effect in another as informative without reporting a test of the interaction between condition or experiment and effect. Similarly, when one correlation or regression coefficient is significant, another is not, and the authors claim that the first coefficient is significantly larger than the second, you can ask for appropriate statistical comparisons to support this claim (Clogg, Petkova, & Haritou, 1995; Steiger, 1980). These nonexhaustive examples illustrate the need for reviewers to be vigilant about appropriate interpretations of nonsignificant results.

Assess constraints on generality

Researchers have always been expected to describe limitations of their research in the Discussion section, but such statements are often pallid, incomplete, and drowned out by louder claims of the importance of the findings. Simons, Shoda, and Lindsay (2017) proposed a stronger and more structured *constraints on generality* (COG) statement, which identifies the aspects of a study (e.g., subjects, materials, procedures, historical and temporal context) that the authors believe are essential to observing the effect. This information is important in evaluating the contribution of a manuscript and for facilitating replications and tests of boundary conditions. Just as important is the fact that the COG statement tends to foster intellectual humility about the generalizability and importance of the finding beyond the limited samples and materials in the research. Some journals already require a COG statement. As a reviewer, you can ask for one as well if the conclusions seem broader than can be justified by the studies.

Writing the Review

Address replicability

An important question to ask yourself when reviewing is, "How confident am I that a direct replication of this study would yield a similar pattern of findings?" Replicability is not the only characteristic of good science—the best work is also interesting, informative, and relevant—but it is a fundamental starting point. We recommend that you cite in your reviews specific reasons why you have (or lack) confidence in the replicability of the work. For example, you may cite statistical robustness, open reporting, and methodological sensitivity as reasons for your confidence in the reported findings.

If replicability is in question, you might suggest that the authors be invited to conduct a preregistered direct replication, perhaps with increased statistical power or other improvements, but designed to replicate the same study as exactly as possible. This invitation may include a no-fault clause making it clear that the new study will be evaluated independently of what the results show, as long as the overall case for the hypothesis is presented reasonably. This approach assumes that similar data can be obtained without tremendous burden (e.g., the methods are not intensive, a convenience sample can be used). If not, you can insist that conclusions be calibrated to the strength of the data. Similarly, openly exploratory work may still be worth publishing if the discussion of results and limitations is appropriate, if the findings are theoretically informed and have

potential to generate new hypotheses, and if the data and materials are publicly available (e.g., McIntosh, 2017).

Communicate your own limits

When you are not familiar with a methodology or statistical test used in a manuscript, it is important to communicate this to the editor, at the same time recognizing that your perspective on other issues may still be valuable. Acknowledging your limits is part of the practice of intellectual humility, and it helps editors become aware when they do have the expertise they need on board. This may lead them to seek out the opinion of an expert in the topic.

Take the right tone

When we asked 22 editors what they would say to reviewers, the most frequent advice was to keep a constructive, respectful tone (see Lindsay et al., 2017). When reviewing with attention to transparency and replicability, it can be tempting to frame departures from best practices as dishonesty or cheating. Indeed, making accusations can be psychologically rewarding (Hoffman, 2014). Not surprisingly, researchers tend to respond defensively when terms like questionable research practices and *p*-hacking are aimed at them. However, many errors happen unintentionally, and many research practices now seen as inappropriate have long been standard in some areas of psychology, entrained by mentors and the gatekeepers of publication. In our view, a polite and reasoned tone is more likely to succeed. Explain the reasons for your recommendations; not all authors or editors are well educated in the new standards. Avoid inflammatory labels in favor of more neutral phrases, such as "low robustness." Always maintain a degree of humility, keeping in mind that your perceptions of flaws may be mistaken.

Promote transparency

If the authors of a manuscript have not followed open-science practices that give reviewers access to materials, analysis code, and data, you may include in your review arguments for making such materials available in subsequent revisions. Your arguments may be directed to the editor as much as to the authors. For example, if the journal endorses the American Psychological Association's ethical standards for publishing, you could ask for a statement of full disclosure of measures, manipulations, and exclusions, because those standards prohibit "omitting troublesome observations from reports to present a more convincing story" (American Psychological Association, 2010, p. 12). To support full disclosure, you could also invoke the American Statistical Association's guideline that *p* values can be interpreted correctly only with full knowledge of the hypotheses tested (Wasserstein & Lazar, 2016) and note that with exploratory analyses, the focus should be on confidence intervals and effect sizes, rather than *p* values. The strongest commitment to openness goals is represented by the Peer Reviewers' Openness Initiative (Morey et al., 2016), whose signatories overtly commit to complete reviews only when all data and materials are made available. No matter what form your request for more openness takes, even if it is denied, it will make the editor and authors more aware of changing norms.

If the authors did provide data, materials, or analysis code, or if they preregistered their research, report in your review what depth of scrutiny you gave to these additional materials. Note any obstacles or limitations you encountered; for example, you might have been unable to check the analysis code because you are not familiar with the programming language used. It is not necessarily your job to make sure those resources are usable and correct. However, reporting the depth of your own efforts will help the editor fulfill his or her obligation.

Some journals offer special recognition in the form of badges granted to articles that meet criteria for transparent processes (e.g., an open-data badge, a preregistration badge, and an open-materials badge; see Blohowiak et al., 2018). If the journal for which you are reviewing offers such badges, consider mentioning that fact, with the aim of encouraging the authors to share more information and improve the review process. If the authors have already applied for one or more badges, keep in mind that most journals rely on authors' declarations that the archived documents are adequate. Authors and readers might benefit from your input if you check badge-supporting material for usefulness and completeness.

Think about signing reviews

Finally, you may also consider breaking the usual anonymity of peer review, signing your reviews to promote transparency and openness on your side of the process. There are good arguments for either signing or not signing all reviews (e.g., Peters & Ceci, 1982, and accompanying peer commentary; Tennant et al., 2017). We recommend adopting a general policy about whether you will or will not sign all reviews, taking into consideration your career stage (see the next paragraph). Without a general policy, you may be tempted to associate yourself with only the reviews that make

a favorable impression (e.g., positive feedback) while avoiding accountability by not signing reviews that make a less favorable impression (e.g., critical feedback). If you do sign, we recommend that you state explicitly that this is a general policy for you, after giving your name.

Signed reviews can have tangible benefits for authors, providing context for suggestions and a sense of fairness in critique, and they give reviewers exposure, credit, and accountability. But signing also carries risk, especially if you are not yet permanently employed. Some authors may seek retribution if they feel their submissions have been inappropriately criticized. Reviewers with more job security and seniority, however defined, have less to lose by signing. These concerns are also relevant when deciding whether to accept requests to review for journals that have adopted open review practices, such as unblinded review, publication of reviews alongside the final article, or direct interaction between authors and reviewers during the review process (see Ross-Hellauer, 2017; Walker & Rocha da Silva, 2015).

Special Cases

Replication studies

The new approach to methods includes a growing willingness to publish reports on close replications of previous research, which previously might have been rejected because they lacked novelty. Main concerns for reviewers are somewhat different for a replication study than for primary research. You do not need to evaluate the theoretical rationale, and your analysis of methods should focus on how closely the replication followed the original, and whether any changes in method were necessary or justified. Brandt et al. (2014) have provided detailed guidance on what makes a replication strong. In brief, just as in the case of original studies, reviewers should give more credence to replications that were preregistered, had adequate power, used methods shown to be sensitive (e.g., manipulations and measures were validated in the new context), and are reported with detailed method sections, open data, and analysis scripts. Given that most journals will publish replication results even if null, it is especially important to reduce the risk that a failure to replicate was due to insensitive methods.

If the authors bill their study as a close (or *direct*) replication, their manuscript should report discrepancies between their study and the original study (Brandt et al., 2014). The importance of these discrepancies depends on the scope of the claims made in the original report. For example, if the samples used in the original and replication studies differed in gender, age, ethnicity, or nationality, you should refer to the original report to assess if its authors generalized their claims across these demographics. If they did, and the replication had weaker or opposite results, it is fair for the replication authors to conclude that their findings reduce confidence in the original claims. However, if the original authors' claims were specific to a population, and the replication sampled a different population, it is not a close replication and does not directly address the original effect. In some cases, discrepancies may need to be introduced in order to reproduce the psychological effect in a new context. For example, if a North American study on perceptions of baseball players is replicated in India, cricket players would be a more appropriate choice to ensure that subjects' knowledge and interest in the material is reproduced.

In reviewing replications, you may have to assess claims about the new state of evidence, taking into account both the original and the replication studies. Gelman (2016) suggested using a time-reversal heuristic to assess the evidence in a replication and the original study: If the replication result had been published first, would it have seemed more compelling than the original result? Just as no single study can determine whether an effect exists, neither can any replication. So, do not be too concerned with judging replications as "successful" or "failed." Instead, think meta-analytically, across the individual studies. Does the replication reinforce or change your beliefs about the effect (or does it do neither)? In any event, it is important to treat positive and negative results in a replication evenhandedly. Although failing to replicate a well-known effect may be more newsworthy than successfully replicating it, both types of evidence need to be reported for science to progress.

Some editors may ask you to judge how important it was to replicate the effect in the first place, just as they would ask you to judge the importance of any novel research. In this case, weigh the strength of existing evidence and the original research's impact on scholarship and society. If the effect has been closely replicated numerous times, has little theoretical or societal value, or has been largely ignored in the academic literature and press, then the replication may be judged as relatively unimportant (Brandt et al., 2014).

Registered Reports

More and more journals are inviting Registered Reports (RRs; see Center for Open Science, n.d.-a) as a special form of preregistered article. Researchers submit a

detailed proposal of a study to a journal for peer review before collecting the data. After data are collected, they submit the complete manuscript reporting results, and the manuscript will be accepted in principle regardless of results if the approved proposal has been followed faithfully. RRs are quite new, but their adoption appears to be increasing rapidly (see Nosek & Lindsay, 2018). Anecdotal reports indicate that reviewers find being involved with RRs gratifying. They can help researchers avoid mistakes in the first place, rather than just pointing out mistakes after they are made.

Peer review of RRs will potentially involve you at two stages. In Stage 1, you will be asked to evaluate the importance and quality of the proposed study prior to data collection. At this stage, evaluate the proposal as you would a normal introduction and Method section, and consider whether the analysis plan makes sense as the complete basis for a Results section. As is true with replications, the possibility of null results means that sensitivity of the methodology is especially important.

After data are collected and analyzed according to the plan, the editor may ask you to assess the report at Stage 2. The manuscript will now have Results and Discussion sections based on the data. At this stage, evaluate whether the research conformed to the plan, whether any changes from the proposal were well justified, and whether other conditions for validity were met (e.g., whether floor and ceiling effects were avoided, the manipulation passed manipulation checks, and the study is accurately and clearly reported). If the answer to these questions is yes, then the manuscript should ultimately be accepted, although revisions might be required to improve readability or to modify the conclusions.

Conclusion

Serving as a peer reviewer provides opportunities to learn about your academic field, to become known and respected (or at least known to and respected by editors), and, most important, to shift norms and shape the future of the field. As best practices in research evolve, so too will best practices in peer review. To contribute to psychology's renaissance (Nelson et al., 2018) and credibility revolution (Vazire, 2018), peer reviewers should promote the good practices of transparency, validity, robustness, and intellectual humility. We hope that these concrete guidelines can help peer reviewers at all career stages provide more effective reviews, and thereby improve the trustworthiness of the published literature and scientific progress as a whole.

Appendix A: Outline of Advice for Promoting Robustness and Transparency When Reviewing Psychology Manuscripts Reporting Quantitative Empirical Research

- Preparing to review
  - Understand what p values mean and do not mean
  - Know the importance of specifying predictions ahead of time
  - Know assumptions underlying frequently used statistical tests
  - Consult the journal's statistical and reporting standards before you review
- Evaluating the reporting of statistics
  - Look for an a priori or sensitivity power analysis (post hoc analyses are not of much use)
  - Look for descriptive statistics, such as means, standard deviations, and correlations
  - Look for a methodological disclosure statement verifying that the article reports all measures, manipulations, and exclusions in the study
  - Consider requesting any of the three preceding elements that are missing
  - Evaluate whether decisions such as analyses, exclusions, and transformations were determined a priori or post hoc, and consider whether more evidence (e.g., replication) may be needed in the case of post hoc analyses
  - Determine whether an optional stopping rule was used in data collection and if it was, how it was corrected for
  - Keep an eye out (e.g., using statcheck) for errors in reporting, such as incorrect degrees of freedom or inferential statistics
  - If you do not know much about some of the techniques used by the authors, acknowledge this to the editor
- Dealing with data, materials, and preregistrations
  - Check the availability of any preregistrations, data, materials, and analysis code
  - Optionally, examine the completeness and accuracy of the available data, materials, and analysis code
  - Consider requesting data, materials, and analysis code if it is missing without good reason
  - Optionally, examine the specificity and completeness of any preregistrations
  - Tell the editor how far you went in checking this material
- Evaluating statistical outcomes
  - Assess the quality of evidence without relying on $p < .05$ per study as either necessary or sufficient for drawing a positive conclusion

  - Assess multistudy reports with the understanding that under complete and transparent reporting, multiple studies all showing *p* values close to .05 are uncommon
  - If you are not sure about the replicability of results, consider requesting a preregistered additional study or more transparent reporting of the existing studies
  - Evaluate claims of null effects as carefully as claims of positive effects (e.g., with Bayesian or equivalence testing)
- Assessing constraints on generality
  - Consider asking for a statement on what aspects of the study the authors believe are essential to observing the effect
- Promoting transparency
  - If the manuscript is not accompanied by shared data, materials, or analysis code, and does not give a good reason for not sharing them, consider requesting them
  - Decide whether you will or will not sign all of your reviews
- Reviewing replications
  - Use the same level of scrutiny for replications as for original studies
  - In the case of a direct replication, assess whether the authors demonstrate that the replication was sufficiently similar to the original study and whether any discrepancies were needed to reproduce the psychological variables in a new context
  - If called upon to examine the need for a replication, consider the strength of existing evidence, the effect's theoretical importance or potential value to society, and the original research's prior impacts on other research and society
  - Do not be too concerned with assessing the success or failure of the replication; think meta-analytically about what the sum of all results says about the effect
- Reviewing Registered Reports
  - Evaluate the proposal's introduction and Method sections as usual
  - Assess whether the analysis plan covers all of the important details and can serve as the complete basis for a Results section
  - For the final report, assess whether the method and analysis plans have been followed faithfully and assess the rationale for any deviations from the proposed plan

Appendix B: Resources on Robustness and Transparency in Psychological Research

This appendix is a list of resources intended to be a useful starting point for reviewers seeking to improve their understanding of the methodological and statistical concepts underlying psychology's credibility revolution. We recognize that there are many more references and resources available; we do not claim that this list is comprehensive or that the resources included represent the "gold standard" among all possible resources.

Open Science

Center for Open Science, https://cos.io/

> The Center for Open Science provides tools, training, support, and advocacy for open-scientific practices. Their Web site contains background on the goals of open science, as well as various services and training opportunities that reviewers can take advantage of to stay up to date with the latest developments.

Open Science Framework, https://osf.io/

> The Open Science Framework (OSF) provides a public repository for researchers to share their materials, data, and analysis scripts. Reviewers can ask authors to consider making the basis of their scientific claims available through the OSF or another public repository.

Center for Open Science. (n.d.-b). [Transparency and Openness Promotion (TOP) guidelines]. Retrieved from https://cos.io/our-services/top-guidelines/

> These eight guidelines (e.g., regarding data transparency) were crafted by a group led by Brian Nosek of the Center for Open Science (Nosek et al., 2015). To date, the guidelines have been implemented (at varying levels of stringency) by 850 journals. Reviewers may want to find out if the journal for which they are reviewing has endorsed the TOP guidelines.

Statistical power

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

> This classic article provides background education on the rationale for power analysis and the sample sizes required for the simplest analyses to have 80% power to detect "small," "medium," and "large" effects.

Magnusson, K. (2018b). Understanding statistical power and significance testing: An interactive visualization [Web app]. Retrieved from http://rpsychologist.com/d3/NHST/

> Reviewers can use this brief primer (with an interactive visualization) to refine their understanding of how power, Type I and Type II errors, effect size, sample size, and alpha are related to each other.

Champely, S. (2018). pwr: Basic functions for power analysis (R package Version 1.2-2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=pwr

This R package provides power-analysis functions that reviewers may want to use to assess the statistical power of reported analyses and may want to recommend to authors if they do not report power analyses. The quick-start guide is available at https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi:10.3758/bf03193146

For reviewers who are not familiar with R, G*Power 3 is another free program with a point-and-click interface that can be used to conduct a range of power analyses during peer review.

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*, 1547–1562 doi:10.1177/0956797617723724

This article provides a readable summary of basic concepts of statistical power. It is similar to other treatments of these concepts but goes beyond them by offering a way to take both publication bias and estimate uncertainty into account when planning sample size. It is useful for evaluating sample-size justifications, especially for replication studies. There is an associated set of apps under the section "Bias and Uncertainty Corrected Sample Size for Power" at https://designingexperiments.com/shiny-r-web-apps/.

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral Anova designs*. Retrieved from https://pdfs.semanticscholar.org/ca52/e5d4976713ecdd62fa10a501d0bf094a30a2.pdf

This power-analysis program provides power calculations for general analysis of variance designs and can flexibly handle designs with any number of fixed or random factors, each with any number of levels, and with any valid pattern of nesting or crossing of the factors. Reviewers might suggest this app to authors in need of power-analysis resources.

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming avoids the term *power* because he believes that psychologists should abandon null-hypothesis significance testing (NHST) in favor of an emphasis on precision of effect-size estimates. But his book is very engaging and compelling in explaining why *p* values are themselves very unreliable. We recommend this resource as background reading for reviewers.

Effect sizes

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. doi:10.1177/0956797613504966

In this article, Cumming encourages researchers to move beyond a focus on statistical significance to an emphasis on effect sizes and confidence intervals. Reviewers may find this article useful for enhancing their understanding of the benefits of such a shift and as a resource to support requests that authors provide confidence intervals and discuss effect sizes.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, Article 863. doi:10.3389/fpsyg.2013.00863

This article is a how-to for navigating the large number of power statistics applicable to designs that compare distinct groups and can inform reviewers' recommendations about power analysis.

Understanding p *values*

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003. doi:10.1037//0003-066x.49.12.997

Cohen reviews the problems with NHST and common misunderstandings about *p* values. Reviewers can read this article to refine their understanding of *p* values.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), Article e124. doi:10.1371/journal.pmed.0020124.

Despite an arguably overstated title, this article makes a compelling case for the limitations of *p* values alone and the need to evaluate truth claims by referring also to statistical power and prior probability. It is useful background reading for understanding the logic of NHST and evidence.

Schönbrodt, F. (2014). When does a significant *p*-value indicate a true effect? Understanding the Positive Predictive Value (PPV) of a *p*-value [Web app]. Retrieved from http://shinyapps.org/apps/PPV/

This interactive demonstration of *p* values' predictive value is based on Ioannidis (2005).

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*, 129–133. doi:10.1080/00031305.2016.1154108

A broad consortium of frequentists and Bayesian statisticians approved this message about the limitations of *p* values, including the need for reporting exact values, additional statistics such as effect sizes, and the full context of analyses. This is a very useful authority that can be cited to support full disclosure and a nuanced approach to significance.

Magnusson, K. (2018a). Distribution of *p*-values when comparing two groups: An interactive visualization [Web app]. Retrieved from http://rpsychologist.com/d3/pdist/

Reviewers can use this interactive app to hone their intuitions about what distributions of *p* values look like under different assumptions.

Sequential analyses

Lakens, D. (2014a). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*, 701–710. doi:10.1002/ejsp.2023

This how-to article argues persuasively that with appropriate reporting and controls, collecting data from subjects in successive groups until a stopping point is reached is not cheating, but rather is an efficient method of collecting data in the face of uncertainty about effect sizes. If it appears that authors sampled sequentially without error correction, reviewers can refer them to this and the following two articles.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*, 293–304. doi:10.1177/1745691614528214

Sagarin et al. present an argument similar to that of Lakens (2014a) and provide a simple method of adjusting *p* values for sequential collection.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322-339. doi:10.1037/met0000061

This article presents a Bayesian approach to sequential testing, which the previous two articles approach using NHST.

Interpreting null results

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*, 328–331. doi:10.1198/000313006X152649

Gelman and Stern explain why differences in statistical significance are often not themselves statistically significant. Reviewers can read this article to become more aware of this issue and cite it as support in reviews.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*, 355–362. doi:10.1177/1948550617697177

In this article, Lakens describes one way to demonstrate evidence for the null within an NHST framework. Reviewers may ask authors to use equivalence tests (or Bayesian methods; see the next section) to provide further context for null findings.

Bayesian approaches

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25*, 5–34. doi:10.31234/osf.io/q46q3

This article explains the probability theory underlying Bayesian analysis and presents some use cases with Harry Potter–themed examples. It is good preparation for evaluating Bayesian analyses, which are becoming more common in submitted manuscripts.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57

Wagenmakers et al. provide 10 arguments for using Bayesian analysis and rebut the most commonly heard objections. They take a somewhat more general approach than Etz and Vandekerckhove (2018) to the goal of understanding the utility of and necessary parameters for Bayesian analysis.

Detecting statistical discrepancies

Nuijten, M. B., Epskamp, S., & Rife, S. C. (2018). statcheck [Web app]. Retrieved from http://statcheck.io/

This app automatically analyzes documents for discrepancies between reported inferential statistics and

p values. Reviewers may wish to run manuscripts through statcheck, either using R or using the online interface.

Preregistration

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research Preregistration 101. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/research-preregistration-101

This article provides an accessible and brief overview (with FAQs) of preregistration. It is a useful introduction to preregistration for reviewers who are unfamiliar with this practice.

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12. doi:10.1016/j.jesp.2016.03.004

This article provides a more extensive discussion about the elements of preregistration, with a proposed standard template.

Open Science Framework. (n.d.). *Registrations*. Retrieved from http://help.osf.io/m/registrations

This guide provides resources and templates for preregistration.

AsPredicted, aspredicted.org

AsPredicted provides a simple framework for preregistration. Reviewers who are new to preregistration might want to consult this template to better understand the key ways in which preregistration can constrain researcher degrees of freedom.

Methodological disclosure and generalizability statements

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Retrieved from https://papers.ssrn.com/abstract=2160588

The simple methodological disclosure statement proposed in this article allows authors to confirm, in 21 words, that they have reported how they determined their sample size, all their manipulations, and all the measures used.

Nosek, B. A., Simonsohn, U., Moore, D. A., Nelson, L. D., Simmons, J. P., Sallans, A., & LeBel, E. P. (2017). *Standard reviewer statement for disclosure of sample, conditions, measures, and exclusions*. Retrieved from https://osf.io/hadz3/

Reviewers can request that authors provide a methodological disclosure statement along the lines of the 21-word solution and recommend this template, which has been endorsed by the Center for Open Science.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*, 1123–1128. doi:10.1177/1745691617708630

Simons et al. propose that authors explicitly define the scope of the conclusions that are justified by the data and specify the target populations (of people, situations, and stimuli) that they expect their findings to be replicable in. Reviewers can ask for such a statement if authors draw conclusions that appear to be broader than justified by the samples used in their manuscript.

Action Editor

Alexa Tullett served as action editor for this article.

Author Contributions

The authors are listed in alphabetical order. All the authors contributed to the generation of the ideas presented in this article and to an initial collaborative draft. Further refinement of this draft proceeded with the input of all the authors, but with D. S. Lindsay and R. Giner-Sorolla doing most of the rewriting.

ORCID iDs

D. Stephen Lindsay https://orcid.org/0000-0002-6439-987X
Jessica P. Lougheed https://orcid.org/0000-0001-8645-9493
Jessie Sun https://orcid.org/0000-0001-6764-0721

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Retrieved from https://www.apa.org/ethics/code/ethics-code-2017.pdf

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*, 1547–1562. doi:10.1177/0956797617723724

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force

report. *American Psychologist*, *73*, 3–25. doi:10.1037/amp0000191

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. doi:10.1038/s41562-017-0189-z

Blohowiak, B. B., Cohoon, J., de-Wit, L., Eich, E., Farach, F. J., Hasselman, F., . . . DeHaven, A. C. (2018). *Badges to acknowledge open practices*. Retrieved from https://osf.io/tvyxz/

Borsboom, D., & Mellenbergh, G. J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. doi:10.1016/j.jesp.2013.10.005

Brown, N. J. L., & Heathers, J. A. J. (2016). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*, 363–369.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*, 550–558. doi:10.1037/a0018933

Center for Open Science. (n.d.-a). *Registered Reports: Peer review before results are known to align scientific values and practices*. Retrieved from https://cos.io/rr/

Center for Open Science. (n.d.-b). [Transparency and Openness Promotion (TOP) guidelines]. Retrieved from https://cos.io/our-services/top-guidelines/

Champely, S. (2018). pwr: Basic functions for power analysis (R package Version 1.2-2). [Computer software]. Retrieved from https://CRAN.R-project.org/package=pwr

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, *100*, 1261–1293. doi:10.1086/230638

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*, 145–153. doi:10.1037/h0045186

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037//0003-066x.49.12.997

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics*. New York, NY: Routledge.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral & Brain Sciences*, *38*, Article e130. doi:10.1017/S0140525X14000430

Epskamp, S., & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute *p* values (R package Version 1.2.2) [Computer software]. Retrieved from http://CRAN.R-project.org/package=statcheck

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. doi:10.31234/osf.io/q46q3

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/bf03193146

Flake, J., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378.

Francis, G. (2014). The frequency of excess success for articles in *Psychological Science*. *Psychonomic Bulletin & Review*, *21*, 1180–1187. doi:10.3758/s13423-014-0601-x

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/measurement-matters

Gelman, A. (2016, January 26). The time-reversal heuristic—a new way to think about a published finding that is followed up by a large, preregistered replication (in context of Amy Cuddy's claims about power pose) [Web log post]. Retrieved from http://andrewgelman.com/2016/01/26/more-power-posing/

Gelman, A., & Loken, P. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331. doi:10.1198/000313006X152649

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social & Personality Psychology Compass*, *10*, 535–549. doi:10.1111/spc3.12267

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*, 200–206.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, Article e1935. doi:10.7717/peerj.1935

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral & Brain Sciences*, *33*, 61–83. doi:10.1017/S0140525X0999152X

Hoffman, M. B. (2014). *The punisher's brain: The evolution of judge and jury*. Cambridge, England: Cambridge University Press.

Howell, D. C. (2010). Statistical methods for psychology (7th ed.). Belmont, CA: Wadsworth.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized

model misspecification. *Psychological Methods*, *3*, 424–435. doi:10.1037/1082-989X.3.4.424

Inzlicht, M. (2015, November). Guest post: A tale of two papers [Web log post]. Retrieved from http://sometimesimwrong.typepad.com/wrong/2015/11/guest-post-a-tale-of-two-papers.html

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. doi:10.1097/EDE.0b013e31818131e7

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*, 196–217.

Kidd, R. F. (1976). Manipulation checks: Advantage or disadvantage? *Representative Research in Social Psychology*, *7*(2), 160–165.

Krueger, J. I., & Heck, P. R. (in press). Putting the p value in its place. *The American Statistician*.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. doi:10.3389/fpsyg.2013.00863

Lakens, D. (2014a). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710. doi:10.1002/ejsp.2023

Lakens, D. (2014b, May 29). The probability of *p*-values as a function of the statistical power of a test [Web log post]. Retrieved from http://daniellakens.blogspot.ca/2014/05/the-probability-of-p-values-as-function.html

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. doi:10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171.

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*, 875–881. doi:10.1177/1948550617693058

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379. doi:10.1037/a0025172

Lindsay, D. S. (2017). *Steve Lindsay on reviewing*. Retrieved from https://osf.io/swgyz

Lindsay, D. S., Giner-Sorolla, R., & Sun, J. (Eds.). (2017). *Digest of tips for reviewers*. Retrieved from https://osf.io/hbyu2/

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research Preregistration 101. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/research-preregistration-101

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585.

Magnusson, K. (2018a). Distribution of *p*-values when comparing two groups: An interactive visualization [Web app]. Retrieved from http://rpsychologist.com/d3/pdist/

Magnusson, K. (2018b). Understanding statistical power and significance testing: An interactive visualization [Web app]. Retrieved from http://rpsychologist.com/d3/NHST/

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*, 2271–2279. doi:10.1080/17470218.2012.711335

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi:10.1146/annurev.psych.59.103006.093735

McIntosh, R. D. (2017). Exploratory reports: A new article type for *Cortex*. *Cortex*, *96*, A1–A4. doi:10.1016/j.cortex.2017.07.014

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., . . . Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547. doi:10.1098/rsos.150547

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. doi:10.1146/annurev-psych-122216-011836

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425.

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Nosek, B. A., Simonsohn, U., Moore, D. A., Nelson, L. D., Simmons, J. P., Sallans, A., & LeBel, E. P. (2017). *Standard reviewer statement for disclosure of sample, conditions, measures, and exclusions*. Retrieved from https://osf.io/hadz3/

Nuijten, M. B., Epskamp, S., & Rife, S. C. (2018). statcheck [Web app]. Retrieved from http://statcheck.io/

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, *48*, 1205–1226. doi:10.3758/s13428-015-0664-2

Open Science Framework. (n.d.). *Registrations*. Retrieved from http://help.osf.io/m/registrations

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral & Brain Sciences*, *5*, 187–195. doi:10.1017/S0140525X00011183

Roediger, H. L., III. (2007). Twelve tips for reviewers. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/twelve-tips-for-reviewers

Ross-Hellauer, T. (2017). What is open peer review? A systematic review [version 2; referees: 4 approved]. *F1000Research*, *6*, Article 588. doi:10.12688/f1000research.11369.2

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293–304. doi:10.1177/1745691614528214

Samuelson, P. L., Jarvinen, M. J., Paulus, T. B., Church, I. M., Hardy, S. A., & Barrett, J. L. (2015). Implicit theories of intellectual virtues and vices: A focus on intellectual humility. *The Journal of Positive Psychology*, *10*, 389–406. doi:10.1080/17439760.2014.967802

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566. doi:10.1037/a0029487

Schönbrodt, F. (2014). When does a significant *p*-value indicate a true effect? Understanding the Positive Predictive Value (PPV) of a *p*-value [Web app]. Retrieved from http://shinyapps.org/apps/PPV/

Schönbrodt, F. D., Maier, M., Heene, M., & Zehetleitner, M. (2015). *Commitment to research transparency*. Retrieved from http://www.researchtransparency.org

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339. doi:10.1037/met0000061

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review*, *2*, 218–226.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Retrieved from https://papers.ssrn.com/abstract=2160588

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128. doi:10.1177/1745691617708630

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. doi:10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2016). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. doi:10.2139/ssrn.2694998

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-*N* design. *Psychonomic Bulletin & Review*. Advance online publication. doi:10.3758/s13423-018-1451-8

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. doi:10.1037/0033-2909.87.2.245

Tennant, J., Dugan, J., Graziotin, D., Jacques, D., Waldner, F., Mietchen, D., . . . Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review [version 3; referees: 2 approved]. *F1000Research*, *6*, Article 1151. doi:10.12688/f1000research.12037.3

Tesser, A., & Martin, L. (2006). Reviewing empirical submissions to journals. In R. J. Sternberg (Ed.), *Reviewing scientific works in psychology* (pp. 3–29). Washington, DC: American Psychological Association.

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. doi:10.1016/j.jesp.2016.03.004

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*, Article 1. doi:10.1525/collabra.74

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*, 411–417.

Veldkamp, C. (2017). Doctoral thesis: The human fallibility of scientists - dealing with error and bias in academic research. *PsyArXiv*. Retrieved from https://psyarxiv.com/g8cjq/

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76. doi:10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Walker, R., & Rocha da Silva, P. (2015). Emerging trends in peer review—a survey. *Frontiers in Neuroscience*, *9*, Article 169. doi:10.3389/fnins.2015.00169

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. doi:10.1080/00031305.2016.1154108

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral Anova designs*. Retrieved from https://pdfs.semanticscholar.org/ca52/e5d4976713ecdd62fa10a501d0bf094a30a2.pdf

Zöllner, S., & Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, *80*, 605–615. doi:10.1086/512821